PCT/IB 03/04400
20. 11. 03

Europäisches Patentamt
European Patent Office
Office européen des brevets

Bescheinigung Certificate Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

02079247.9

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Anmeldung Nr:
Application no.: 02079247.9
Demande no:

Anmeldetag:
Date of filing: 09.10.02
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

Koninklijke Philips Electronics N.V.
Groenewoudseweg 1
5621 BA Eindhoven
PAYS-BAS

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

Security camera video authentication

In Anspruch genommene Priorität(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/
Classification internationale des brevets:

G06T1/00

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL PT SE SK TR

SECURITY CAMERA VIDEO AUTHENTICATION

D.K. Roberts

Philips Research, Prof. Holstlaan 4, 5656 AA, Eindhoven, The Netherlands
keith.roberts@philips.com, Tel: +31 40 27 45264, Fax: +31 40 27 42566

ABSTRACT

The ability to authenticate images captured by a security camera, and localise any tampered areas, will increase the value of these images as evidence in a court of law. This paper outlines the challenges in security camera video authentication, and discusses the reasons why *fingerprinting*, a robust type of digital signature, provides a solution preferable to semi-fragile watermarking. A fingerprint based upon block DC differences is described, and the tamper detection performance is assessed both theoretically and experimentally, the latter using an automated method of 'tampering' images.

1. INTRODUCTION

The need for image authentication arises from the ease with which digital images may be edited and altered. Appropriate authentication techniques should be capable of verifying that the image content is unchanged, and of localising any image portions that have been altered.

A difficulty, however, is that certain changes to the images are allowable, and should not be classified as malicious tampering. For example, lossy compression causes image modifications, but not to an extent that alters a person's interpretation of the image. These modifications must not be mistaken for tampering.

The issues surrounding image authentication are discussed in this paper within the context of security cameras. This application aims to increase the value of video evidence in a court of law by demonstrating that tampering of the video has not taken place.

Section 2 sets the scenario for security camera video authentication. This is followed in Section 3 by a summary of the alternative authentication methods of semi-fragile watermarking and digital signatures. Section 4 describes a solution appropriate for security cameras, and the performance of this system is presented in Section 5.

2. AUTHENTICATION REQUIREMENTS

Figure 1 illustrates the layout of a typical surveillance system. This consists of the following components:

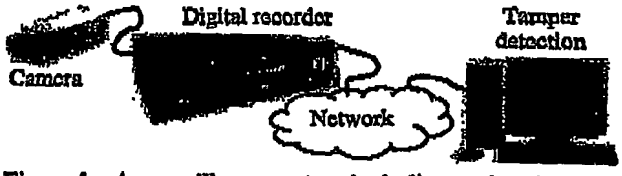

Figure 1 - A surveillance system including authentication

1. Camera – the video output is usually in an analogue format, such as PAL or NTSC
2. Digital recorder – this takes the video inputs from multiple cameras and applies lossy compression
3. Computer network – this provides storage, retrieval, and authentication means for the compressed video

2.2. Allowable processing

Figure 1 shows the allowable image alterations: between capture and authentication the video travels through an analogue link, and is then compressed. The distortions to the video that result must not be mistaken for tampering.

A variety of compression methods are in use in surveillance systems, including both spatio-temporal (e.g. MPEG), and still-image techniques (e.g. JPEG, ADV601 [1]). Where still-image compression is applied, compression in the temporal direction is achieved by retaining, for example, only one image every 5 seconds.

2.3. Tampering

The envisaged type of tampering is pixel replacement. For example, this could be the removal of a person by replacement with 'background' content, perhaps copied from an earlier/later image in which the person is absent.

A guideline for the minimum detectable size of tampered region is the minimum size at which a human face is recognisable. The information in [2] points to a size of approximately 33 pixels wide and 50 pixels high for PAL/NTSC video content.

3. ALTERNATIVE APPROACHES

Tamper detection proceeds by comparing authentication data derived from the suspect image with the corresponding data derived from the original image. This

can be decomposed into two sub-problems:

1. How to generate appropriate authentication data
2. How to transport the authentication data of the original image to the point in the system where authenticity is tested

At the camera it is not known whether the recorder will discard images during compression. The authentication data must therefore be generated and transported such that each image may be authenticated independently, without reference to images at any other point in time.

The ability to distinguish between allowable and malicious alterations is usually referred to by the term *semi-fragile*. There are two alternative authentication solutions depending upon where this fragility is located:

1. Semi-fragile watermarks: the *transport* of the original image's authentication data is such that it can be correctly retrieved after allowable alterations, but not after tampering
2. Semi-fragile digital signatures: the *generation* of the authentication data is such that the data is invariant to allowable alterations, but not to tampering

These two alternative approaches will now be discussed in more detail.

3.1. Semi-fragile watermarking

Semi-fragile watermarking [3,4] usually generates a fixed pattern of bits for the authentication data, and then embeds these using a semi-fragile technique. Authenticity checking consists of extracting the watermark bits and comparing them against the pattern that was embedded. The locality of tampered image regions is indicated by errors in the extracted authentication bits.

The use of a fixed pattern of embedded bits facilitates the creation of apparently authentic tampered images. For example, pixels can be replaced by content copied from the same location in a different, but authentic, image [5]. Extraction of the watermark bits will still be successful, and so the altered image will be judged authentic.

Security can be increased by generating the authentication bits such that they are dependent upon the image content. This can help prevent the copy attack example given above. If the content dependent watermark bits also possess fragility to tampering, then such a scheme has properties of both semi-fragile watermarking and semi-fragile signatures. If, for example, the authentication data and watermark are fragile to different types of image alterations, then this approach can help indicate what type of tampering has taken place.

Semi-fragile watermarking can only protect the image features (e.g. pixels or frequency coefficients) that are used for embedding the authentication data. Protecting the most perceptually important image features therefore requires data to be embedded into these features. This may present difficulties in ensuring watermark invisibility.

Any image material in which watermark bits cannot be both invisibly embedded and reliably detected (e.g. flat content), will result in bit errors even without tampering [6]. There is no way to distinguish these bit errors due to zero watermark capacity from those due to tampering. The replacement of original image regions by flat content can therefore create an apparently authentic tampered image.

In [7] an attempt is made to overcome this problem via 'backup embedding'. Each watermark bit is embedded twice, using two spatially separate embedding locations. However, in the presented method there is no guarantee that the backup location does not also have zero watermark capacity. Embedding each authentication bit multiple times must also have negative implications for either the tamper localisation ability (due to fewer authentication bits for a given embedding capacity), or for invisibility and robustness to allowable operations (due to an increased number of embedded bits).

3.2. Semi-fragile digital signatures and fingerprints

A digital signature is a set of authentication bits that summarise the image content. A semi-fragile signature is generated in such a way that a tampered image gives a changed set of summary bits, but an image processed only by allowable manipulations does not. This non bit-sensitive type of signature will be referred to as a *fingerprint* in order to provide a clear distinction from cryptographic digital signatures, and highlight the relevance to other applications [8].

The image features from which the fingerprint bits are calculated may be chosen to give the most appropriate trade-off between robustness to allowable processing, fragility to tampering, and computational cost. There are many suggestions in the literature including: DC values [8], moments [9], edges [9], histograms [10], compression invariants [11], and projections onto noise patterns [12].

Authenticity is verified by comparing the fingerprint generated from the suspect image, with the original fingerprint calculated in the camera. Typically, a direct relationship exists between individual fingerprint bits and an image location. For example, the image may be split into blocks and a bit derived for each block. The locality of tampered image regions is therefore indicated by which particular fingerprint bits are in error.

Note that there is a trade-off between the number of fingerprint bits and the localisation ability. For example, a smaller block size allows better localisation of tampered areas, but there are more blocks per image, and thus more fingerprint bits.

Having generated a fingerprint of the original image in the camera, there remains the problem of transporting this fingerprint data, such that it is available at authenticity verification. One possibility is to embed the fingerprint bits into the image as a watermark. This, and alternatives

for handling the fingerprint as 'meta-data', are discussed in the following two sub-sections.

3.2.1 Transport by watermark

Watermarking provides a neat solution to the transport problem: by invisibly embedding the fingerprint into the image, this data is automatically carried with the image. Clearly the watermark must be robust to (at least) all allowable image processing. If the watermark is also semi-fragile, this can aid identification of the type of tampering that has occurred, as explained in Section 3.1. The content dependent nature of the fingerprint bits also helps prevent watermarked content copied from one image to another from appearing authentic.

A fingerprint protects against alteration of the image features used to calculate the fingerprint bits. These features may be different from those used to embed the fingerprint as a watermark. This gives increased flexibility to embed bits in the most appropriate manner for invisibility and robustness requirements, and helps avoid the zero watermark capacity problems from which semi-fragile watermarking authentication schemes suffer.

A drawback of transporting fingerprint data using a watermark is that this may limit the tamper localisation ability. A sufficiently robust watermark will typically have a fairly limited payload size, which may place an unacceptable constraint upon the fingerprint size, and hence upon the localisation ability.

3.2.2 Transport by meta-data

Transporting fingerprint data separate from the video is not possible due to the analogue cable between the camera and recorder. This requires that the authentication data generated in the camera must be embedded into the video signal itself for transmission to the recorder.

An alternative to watermarking, is to embed the fingerprint data directly into the pixel values, in a manner similar to teletext data in television signals. Security cameras currently transport camera parameters, control information, and audio using such data channels. The data carrying capacity of these data channels can be far greater than a watermark, depending upon how many video lines are utilised. If only video lines in the over-scan area (vertical blanking interval) are employed, then invisibility of the embedded data is maintained.

It is important that fingerprint data is encrypted before it is embedded in this manner. Without encryption, substitution of the original fingerprint data with a fingerprint corresponding to a tampered image would make the forgery appear authentic. Missing or damaged authentication data must always be interpreted as tampering.

4. A SOLUTION FOR SECURITY CAMERA VIDEO AUTHENTICATION

An authentication system for security cameras is currently in development. The chosen approach embeds fingerprints into the over-scan area of the video signal. The details of the fingerprint formation and authentication are presented in the following sections.

4.1. Fingerprint calculation

Fingerprints should be calculated based upon the low frequency content of the image. This is necessary to provide resilience to the analogue link, which severely limits the video signal bandwidth, and lossy compression, which typically discards the higher frequency components.

In applications where the allowable processing operations are well characterised, this knowledge may be utilised in fingerprint calculation. For example, properties that are invariant to JPEG quantisation are used to form fingerprints in [11]. However, due to the wide variety of compression methods used in surveillance systems (see Section 2.2) such an approach is not possible.

As mentioned in Section 3, the camera must calculate and embed authentication data in real-time for each and every output image. This places severe constraints upon the computational load if the impact upon the camera cost is to be minimised.

A low frequency and low complexity fingerprint can be formed by utilising only the DC component. The image is divided into blocks, and differences between blocks' DC values (mean pixel luminance) are used to form the fingerprint. Using DC differences provides invariance to changes in the overall image DC component, e.g. due to brightness alterations. Taking differences between the DC values of adjacent blocks captures how the image content of each block relates to its neighbours. More specifically, a fingerprint bit b_i is derived for the i^{th} block as follows:

$$s_i = \sum_{j=1}^{8} (DC_i - DC_j) \qquad (1)$$

$$b_i = 1 \quad \text{if} \quad s_i > 0, \quad b_i = 0 \quad \text{otherwise}$$

where j indexes the eight blocks that neighbour block i.

The appropriate block size is related to the size of image feature upon which tamper detection is desired. Smaller blocks increase the likelihood of alterations being detected, but at the cost of an increased number of fingerprint bits to calculate and transport.

4.2. Authenticity verification

The most straight-forward approach to checking authenticity is a simple bit by bit comparison of the original and suspect authentication bits. This alone,

however, is unlikely to be satisfactory, as some bit errors due to allowable processing are almost inevitable [4].

Methods to solve this problem are often based upon the observation that these bit errors due to allowable processing are likely to be lightly distributed over the whole image, whereas bit errors due to tampering are likely to be concentrated in a confined area. Allowable operations can therefore be distinguished from tampering via a post-processing operation upon the bit errors, such as error relaxation [9], or mathematical morphology [13].

Note that authenticity verification can afford more complex computation than fingerprint calculation, as it will occur relatively infrequently, need not be real-time, and has a more powerful computation platform available.

Rather than applying an 'after-thought' post-processing step to provide resilience to allowable processing, it is preferable to build this robustness more closely into the authenticity decision. This can be achieved by using 'soft-decision' information during comparison of the suspect image's fingerprint with the original fingerprint bits. This prevents tampering from being indicated in cases where s_i is close to zero, and therefore a fingerprint bit error is likely to occur due to allowable processing.

The authenticity decision for an individual block can be expressed as a choice between hypothesis H_0 (the block's image content is authentic), and hypothesis H_1 (the block's image content has been tampered with). Given the value s of the block (computed according to Equation 1), and the fingerprint bit of the original image b_{orig}, the hypothesis with the greatest probability is chosen:

If $\Pr[H_0 \mid b_{orig}, s] > \Pr[H_1 \mid b_{orig}, s]$, choose H_0

but, from Bayes theorem:

$$\Pr[H_0 \mid b_{orig}, s] = \frac{p_{S|H_0,b_{orig}}(s)\Pr[H_0]}{p_S(s)}$$

and similarly for H_1, so the decision rule becomes:

$$\text{If} \quad \frac{p_{S|H_0,b_{orig}}(s)}{p_{S|H_1,b_{orig}}(s)} > \frac{\Pr[H_1]}{\Pr[H_0]}, \quad \text{choose } H_0 \qquad (2)$$

It is difficult to assign values to the prior probabilities of each hypothesis (this would be equivalent to stating what proportion of images are tampered), so the Neyman-Pearson [14] decision rule is more appropriate. This approach maximises the probability of tampering being detected for a fixed 'false alarm' probability of allowable processing being mistaken for tampering. In practice this simply results in the priors being replaced by a threshold λ, which is set to achieve the desired false alarm rate:

$$\text{If} \quad \frac{p_{S|H_0,b_{orig}}(s)}{p_{S|H_1,b_{orig}}(s)} > \lambda, \quad \text{choose } H_0 \qquad (3)$$

If hypothesis H_1 is true, then we have no knowledge of the replacement content and can only assume that the result of Equation 1 is distributed as for image content in general, i.e. $p_{S|H_1,b_{orig}}(s) = p_S(s)$.

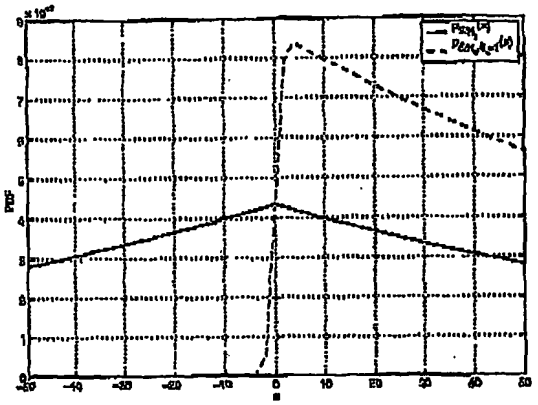

Figure 2 - Conditional PDFs of S under each hypothesis

The probability density function (PDF) $p_S(s)$ has been estimated from a set of images, and turns out to be well approximated by a laplacian distribution. Arguments similar to those in [15] can be used to explain this shape, but are omitted here.

If hypothesis H_0 is true, then the outcome of Equation 1 for the original image, S_{orig}, is of known sign, given by the value of b_{orig}. The distribution of S_{orig} is therefore the one-sided version of $p_S(s)$ (i.e. exponential). Allowable processing operations then cause an error E, resulting in the observed value $S=S_{orig}+E$. The distribution of E should be estimated for the harshest allowable processing to which images will be subject, e.g. the lowest JPEG quality factor. Typically a gaussian distribution provides a reasonable approximation to the PDF of E. Finally, assuming independence of S_{orig} and E, the following convolution gives the PDF required for the hypothesis test:

$$p_{S|H_0,b_{orig}}(s) = \int_{-\infty}^{\infty} p_{S_{orig}}(s-e)\,p_E(e)\,de$$

Figure 2 shows a plot of this PDF for the case of E corresponding to JPEG compression of quality factor 50, and $b_{orig}=1$. Note the deviation from the exponential shape, which is due to E. This gives non zero probabilities of S being negative, and thereby models fingerprint bit errors due to allowable processing.

From Figure 2, whatever the value of the threshold λ, the PDFs only cross at a single point. The hypothesis test therefore reduces to a simple threshold test on blocks' values of S. The threshold value s_T for $b_{orig}=1$ satisfies:

$$p_{S|H_0,b_{orig}=1}(s_T) = \lambda p_{S|H_1}(s_T)$$

and, by symmetry, the threshold for $b_{orig}=0$ is $-s_T$.

It is also clear from Figure 2 that a feature S possessing a less peaked PDF is desirable. This would reduce the smearing over the bit threshold due to E, giving fewer fingerprint bit errors due to allowable processing.

Note that the above derivations assume that values of S are independent and identically distributed for different

blocks. In practice this is not quite true, and some correlation exists between values of S for adjacent blocks. Nevertheless, as will be seen in the results of Section 5, the approach appears very useful.

An advantage of the above hypothesis test framework is that it allows the possibility of errors in the original fingerprint bits to be taken into account. This is achieved by making the value of b_{orig} a random variable distributed according to the bit error rate of the transport channel.

A further advantage of the presented approach is that improvements in the localisation of tampered areas are possible by adjusting the operating point (i.e. the threshold λ). Normally λ is set to achieve the desired low false alarm rate. However, once one or more blocks are identified as tampered, the image as a whole is known to be inauthentic, and each individual block may be considered equally likely to be tampered or authentic. This points towards re-evaluating the authenticity decision for all blocks using equal prior probabilities (i.e. λ=1). This approach can be taken even further by taking the spatial distribution of tampered blocks into account. For example, a block with several tampered neighbouring blocks is also likely to be tampered. These beliefs can be expressed by modifying the prior probabilities, or equivalently, the value of λ. Experiments have shown that these adjustments of the operating point and re-evaluation of authenticity decisions can help extract the size and shape of the tampered region with greater accuracy.

5. RESULTS

The performance of an authentication system can be measured by its probability of detecting tampering, and its false alarm probability when only allowable image processing has been applied. Few publications provide this information, usually giving only one example image on which the authentication method is demonstrated. The detection probability in particular is difficult to assess as it requires the tampering of a large number of images, and manually replacing sections of an image in a convincing way is very time consuming.

To overcome this, the detection rate has been estimated by an automatic process that blends image content from a second unrelated image into the image under test. Many trials are performed, using different test images, different tampered locations, and different replacement image content. The whole test is also repeated for different sizes of tampered area in order to gain a full picture of the performance of the authentication method.

Figure 3 shows an example image and a version of it that has been tampered under the above procedure. A block of nxn pixels has been replaced, of which the central $\frac{n}{2} \times \frac{n}{2}$ pixel block is entirely new content, and the surrounding pixels are a blend of the original and replacement content formed using a window function.

Figure 3 - Example 'tampered' image

The measured false alarm and detection probabilities using this 'simulated tampering' are given in Figures 4 and 5 as a function of the decision threshold s_T. The presented results are for a fingerprint of 1 bit per 32x32 block of pixels, and allowable processing of JPEG quality factor 50. Figure 4 shows that the false alarm probability exhibits the expected transition around the fingerprint bit threshold of $S=0$. The sharpness of the transition is due to the high robustness of the property S to JPEG compression, and consequently small chance of allowable processing causing fingerprint bit errors. Figure 5 shows the detection probability for two different sizes of tampered area. It is clear that for good detection rates, the fingerprint block size is required to be smaller than the minimum size of tampered area that it is wished to detect.

The performance of the authentication system can also be estimated theoretically using the probability distributions derived in the previous section. The detection and false alarm probabilities for an individual block are:

$$\Pr(D) = \int_{-\infty}^{s_T} p_{S|\overline{P}_1}(s)ds = \int_{-s_T}^{\infty} p_{S|P_1}(s)ds$$

$$\Pr(FA) = \int_{-\infty}^{s_T} p_{S|H_0 b_{orig}=1}(s)ds = \int_{-s_T}^{\infty} p_{S|H_0 b_{orig}=0}(s)ds$$

Assuming the individual block decisions to be independent, the false alarm probability for the entire image can be estimated as:

$$\Pr(\text{False Alarm}) = 1-(1-\Pr(FA))^N$$

where N is the number of fingerprint blocks in the image. This is plotted in Figure 4 and can be seen to show good correspondence with the experimental results. This justifies using the theoretical approach to calculate the value of s_T to be used in practice, where a false alarm rate too low to be simulated in a reasonable time is required.

The detection probability for the whole image can similarly be estimated by:

$$\Pr(\text{Detection}) = 1-(1-\Pr(D))^M$$

However, setting the value of M, the number of tampered blocks, is problematic as it is dependent upon the size and

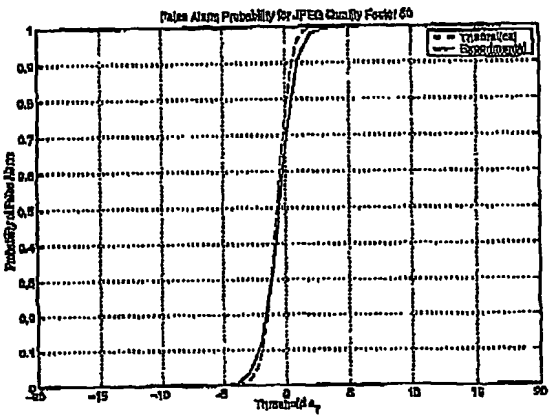

Figure 4 - Probability of false alarm

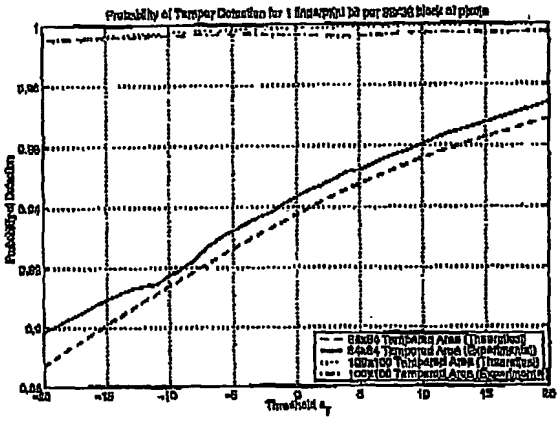

Figure 5 - Probability of tamper detection

shape of the tampered region with respect to the fingerprint blocks. In Figure 5 the detection probabilities are estimated by setting:

$$M = \frac{n^2}{b^2}$$

where the tampered area is a block of *nxn* pixels, and the fingerprint is formed using blocks of *bxb* pixels. This can be seen to give a reasonable match to the experimental results, and is thus a useful estimation of the detection rate when setting the decision threshold.

6. CONCLUSIONS

A fingerprinting solution for security camera video authentication has been presented. Fingerprints based upon block DC differences have been shown to give a good trade between compression robustness, sensitivity to tampering, and computational cost.

A hypothesis test approach to authenticity verification has been introduced. This offers advantages of: tolerance to fingerprint bit errors caused by allowable processing; the ability to cope with bit errors in the received original fingerprint; and improved localisation of tampering by adjustment of the prior probabilities.

7. REFERENCES

[1] www.analogdevices.com

[2] European Standard EN50132-7: "Alarm Systems & CCTV Surveillance Systems Application Guidelines"

[3] I.J. Cox, M.L. Miller, J.A. Bloom, "Digital Watermarking", Morgan Kaufman Publishers, 2002

[4] O. Ekici, B. Coskun, U. Naci, B. Sankur, "Comparative Assessment of Semi-Fragile Watermarking Techniques", SPIE Multimedia Systems and Applications, August 2001

[5] J. Fridrich, "Methods for Tamper Detection in Digital Images", Proc. ACM Workshop on Multimedia and Security, Orlando, October 1999

[6] J.J. Eggers, B. Girod, "Blind Watermarking Applied to Image Authentication", IC-ASSP, Salt Lake City, May 2001

[7] M. Wu, B. Liu, "Watermarking for Image Authentication", Proc. ICIP '98, Chicago, October 1998

[8] J.C. Oostveen, A.A.C. Kalker, J.A. Haitsma, "Feature Extraction and a Database Strategy for Video Fingerprinting", 5th International Conference on Visual Information Systems, Taipei, March 2002

[9] M.P. Queluz, "Content Based Integrity Protection of Digital Images", SPIE Conference on Security and Watermarking of Multimedia Contents, San Jose, January 1999

[10] M. Schneider, S. Chang, "A Robust Content Based Digital Signature for Image Authentication", Proc. ICIP '96, Lausanne, Switzerland, October 1996

[11] C-Y. Lin, S-F. Chang, "Generating Robust Digital Signature for Image/Video Authentication", ACM Multimedia and Security Workshop, Bristol, England, September 1998

[12] J. Fridrich, "Image Watermarking for Tamper Detection", Proc. ICIP '98, Chicago, October 1998

[13] F. Bartolini, A. Tefas, M. Barni, I. Pitas, "Image Authentication Techniques for Surveillance Applications", Proc. of IEEE, September 2001

[14] J.D. Gibson, J.L. Melsa, "Introduction to Non-Parametric Detection with Applications", IEEE Press, 1996

[15] E.Y. Lam, J.W. Goodman, "A Mathematical Analysis of the DCT Coefficient Distributions for Images", IEEE Trans. on Image Processing, October 2000

Improved Localisation of Image Tampering

Introduction

The ease with which images and video may be edited and altered when in digital form stimulates the need for means to be able to authenticate content as original and unchanged. Where it is judged that an image has been altered, it is also desirable to have an indication of which image areas have been changed.

The authentication problem is complicated by the fact that some image alterations are acceptable, such as those caused by lossy compression. These changes may cause slight degradation of the image quality, but do not affect the interpretation or intended use of the image. The result is that classical authentication techniques from cryptography are not appropriate, as typically these methods would interpret a change of just one bit of an image as tampering.

There are two approaches for robust (i.e. not bit sensitive) image authentication that appear in the literature: semi-fragile watermarking, and robust digital signatures (also known as 'fingerprints'). Both of these approaches basically boil down to making a comparison between a set of bits calculated from the suspect image and the corresponding set of bits calculated from the original image content. The descriptions and invention that follow are applicable irrespective of whether these authentication bits constitute a watermark or a fingerprint.

Background

Authentication bits are derived from the suspect image, by computing some property, S, of the image pixel values, and then thresholding S to give either a '0' or '1' bit. The computed property depends upon the watermarking or fingerprinting scheme being used. Typically, an image will be divided into blocks (of, for example, 16x16 pixels), and an authentication bit is generated for each block. This allows localisation of image alterations, as an error in a particular bit can be related to an alteration of a particular image region.

For each of the original authentication bits, a decision must be made whether the suspect image is likely to generate a matching authentication bit or not. This equates to judging whether the corresponding image block is authentic or altered. If a block is judged to be tampered, and the image content has indeed been altered, this is called a *detection*. If, on the other hand, a block is judged tampered when in fact its content has only undergone allowable operations (e.g. compression), the decision is incorrect, and is called a *false alarm*.

A crude system makes the authentication decision by comparing the bits derived from the suspect image against the original authentication bits. A more sophisticated approach is to use 'soft decision' information. In this case the unthresholded values of the property S calculated from the suspect image are used to judge authenticity. Values of S that are on the wrong side of the threshold to generate a bit matching the original authentication bit may still be judged authentic if they are close to the threshold. This gives more robustness to allowable image operations, reducing the probability of false alarms occurring.

Setting exactly which range of values of S will be classified as authentic, and which as tampered, fixes the false alarm and detection probabilities. According to where the decision boundary is placed, different trade-offs between the detection and false alarm probabilities can be achieved. This is often displayed in a Receiver Operating Characteristic (ROC). A typical shape of ROC curve is displayed in Figure 1.

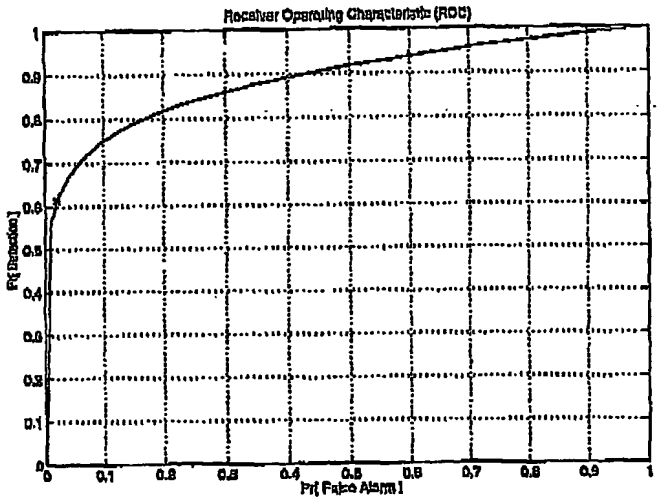

Figure 1 – Example ROC curve

In image authentication, it is expected that only a small minority of images will actually be tampered. It is therefore important to have a low probability of false alarm, otherwise large numbers of authentic images will be declared tampered. The operating point on the ROC curve will therefore usually be chosen to give an acceptably small false alarm rate.

Selecting an operating point that gives a low false alarm probability also reduces the detection probability, as illustrated in Figure 1. This means that many tampered blocks will not be detected. Assuming that the tampered region spans multiple authentication blocks, then the probability of *all* of the altered blocks not being detected is much smaller, so the fact that the image is inauthentic will still be apparent.

Although a low false alarm operating point can still achieve a good probability of detecting whether images have been altered, it has more serious implications for the localisation of image alterations. The low detection probability for individual blocks leads to a patchy detection of which image regions have been changed. This is illustrated in the Figures that follow: Figure 2 shows the original image, and Figure 3 the altered version; Figure 4 shows which authentication blocks are judged as tampered.

It can be seen in Figure 4 that numerous image blocks are judged as tampered, so it is clear that the image is inauthentic. However, comparison between Figures 2, 3, and 4 illustrates the patchy detection of the tampered image area; the full size and shape of the altered image region is not readily apparent.

The object of the invention presented in the next section is to improve the localisation of altered image regions.

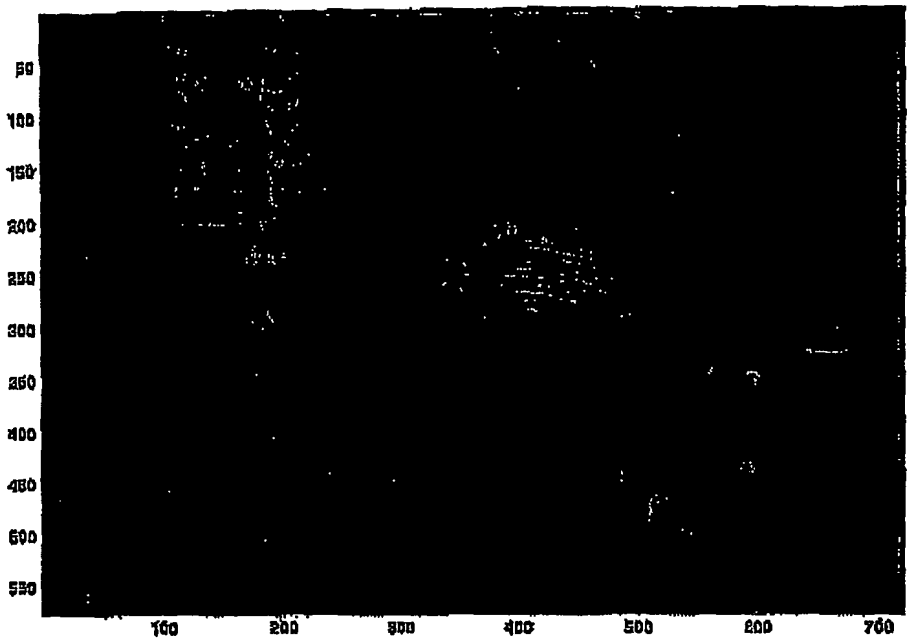

Figure 2 – Original image

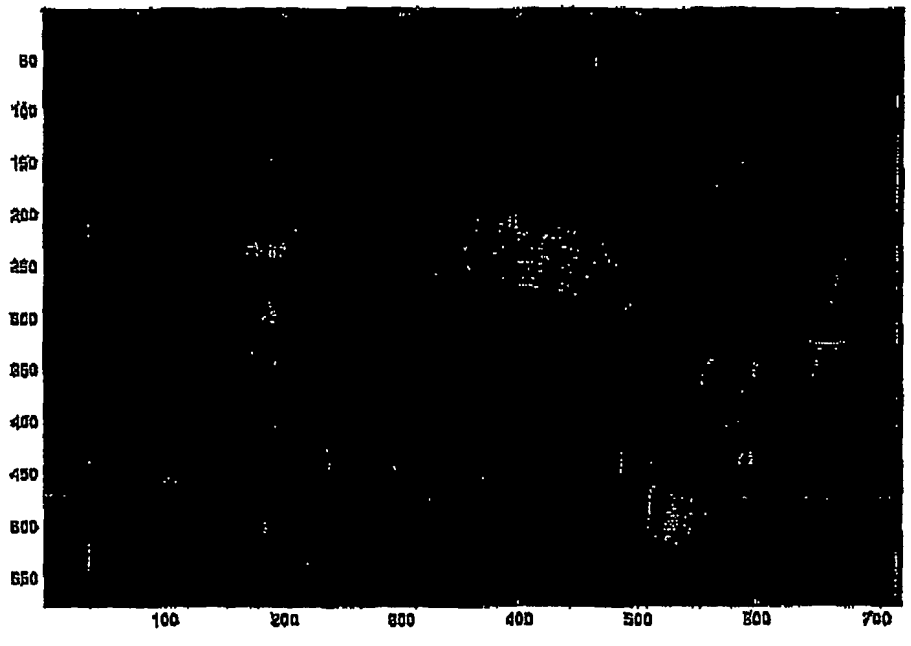

Figure 3 – Tampered image

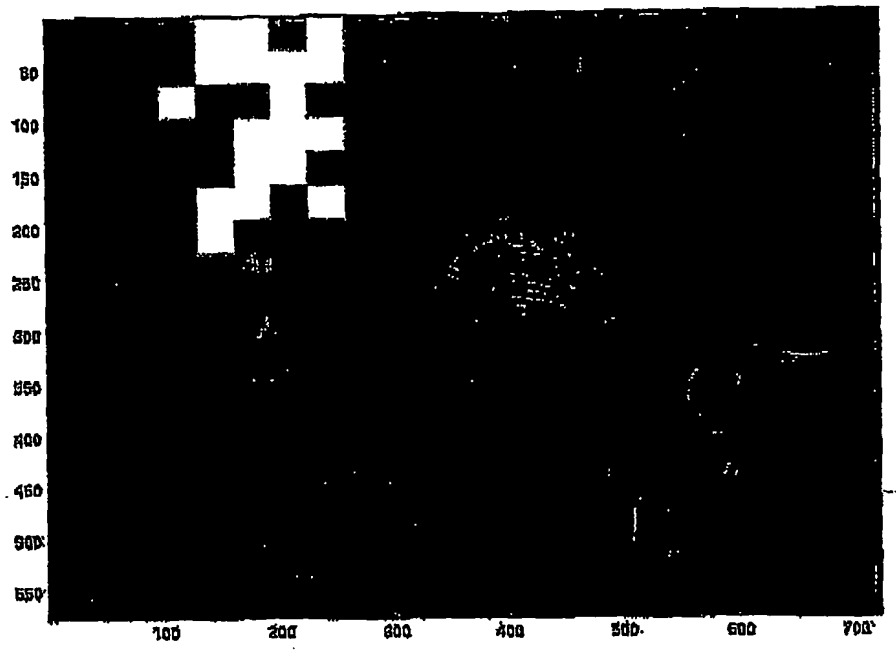

Figure 4 – Tamper detection results

The Invention

The idea is to utilise 'context' information in the authentication decision for each block. In other words, the number and location of blocks that are declared tampered affects the decisions about which other blocks may be tampered. For example, blocks neighbouring a tampered block are under greater suspicion than blocks further away. This context information can be incorporated into the authentication decisions by adjustments to the operating point on the ROC curve.

The authentication check for an image therefore proceeds as follows:

1. An authentication decision is made for each block independently using a low false alarm operating point
2. If no blocks are declared tampered, then the image is taken as authentic
3. If one or more tampered blocks are found then it is known that the image as a whole is inauthentic. This means that blocks neighbouring those that are tampered are also likely to be tampered, and all other image blocks can be assumed equally likely to be authentic or tampered. Knowing this, new operating points are selected for each block's authentication decision.
4. The authentication decisions for all blocks not (yet) declared tampered are re-evaluated using the new decision boundaries
5. If further blocks are declared tampered, the procedure of adjusting the decision boundaries and re-evaluating blocks' authenticity is repeated. This continues until no further tampered blocks are identified

Alterations to the decision boundary can be used to move the operating point to a position with a larger detection probability. This can find further tampered blocks, and thus help determine the full size and shape of the tampered image region. Applying this procedure to the example shown in Figure 4, provides the result shown in Figure 5. The much fuller coverage and localisation of the tampered region is evident.

Figure 5 – Tamper detection results using the invention

Note that the concept of adjusting the operating point on the ROC curve, and re-evaluating decisions in the light of neighbouring decisions, is of value not only in image or video or audio authentication, but is equally applicable to other fields where many inter-related decisions have to be taken.

Example Implementation

To further elucidate the invention, an example implementation will be given. First some background concerning decision theory is required.

One framework for making the authentication decision for each block is to use a hypothesis test. Given the value of the property S calculated for the suspect image block, the hypothesis that the block is tampered (H_1) is selected if this has a greater probability than the hypothesis that the block is authentic (H_0):

$$\text{Select } H_1 \text{ if: } \Pr(H_1 / S = s) > \Pr(H_0 / S = s)$$

Expanding this in terms of the probability density functions of S, and the prior probabilities of each hypothesis gives:

$$\text{Select } H_1 \text{ if: } \frac{p(s / H_1)\Pr(H_1)}{p(s)} > \frac{p(s / H_0)\Pr(H_0)}{p(s)}$$

Rearranging:

$$\text{Select } H_1 \text{ if: } \frac{p(s / H_1)}{p(s / H_0)} > \frac{\Pr(H_0)}{\Pr(H_1)}$$

The difficulty with this decision process is setting the values of the prior probabilities, $\Pr(H_1)$ (the probability that any given image is tampered), and $\Pr(H_0)$ (the probability that any given image is authentic). These probabilities are unlikely to be known, so instead their ratio can be represented by a value λ:

$$\text{Select } H_1 \text{ if: } \frac{p(s / H_1)}{p(s / H_0)} > \lambda$$

The decision process can now be seen as comparing the likelihood of the value s being generated by altered image content, against the likelihood of it being generated by authentic content. The decision boundary is determined by the value of λ. Different values of λ result in different false alarm and detection probabilities, allowing a ROC curve to be plotted. Choosing a value for λ to give a specific false alarm probability therefore selects the operating point on the ROC curve. This approach is known as the Neyman-Pearson decision criterion, and can be shown to maximise the detection probability for a chosen probability of false alarm.

Using this decision framework, the invention can be applied as follows:

1. An operating point λ_0 is chosen that gives an acceptably low false alarm rate. The authenticity of all image blocks is assessed using this decision threshold
2. If no blocks are declared tampered, then the image is taken as authentic
3. If one or more tampered blocks are found, then for all other blocks i, a new operating point λ_i is determined. This adjustment of the decision threshold will take into account the number of tampered blocks found, and their proximity to the block i. Many algorithms for adjusting the decision threshold are possible. One example is:

   $$\lambda_i = \alpha\lambda_1 + (1-\alpha)\lambda_2$$

   where $\lambda_1 = 1$ (representing equal prior probabilities), $\lambda_2 > 1$ (giving a higher detection probability), and α is given by:

   $$\alpha = \left(\frac{n}{8}\right)\left(\frac{d - r_m}{d-1}\right) \qquad \text{and} \qquad r_m = \min(r, d)$$

   where n is the number of the 8 blocks neighbouring block i that are marked as tampered, r is the distance (in units of blocks) of block i from the closest tampered block, and d is some maximum distance that sets how widely around a tampered block that suspicion is raised
4. The authentication decisions are re-evaluated using the new decision boundaries λ_i
5. If further blocks are declared tampered, the procedure of adjusting the decision boundaries and re-evaluating blocks' authenticity is repeated. This continues until no further tampered blocks are identified

This example makes it clear that adjusting the operating point is equivalent to adjusting the prior probability of a block being tampered. This is turn is justified by the block's context, i.e. its location with respect to other tampered areas.

CLAIMS

1. A method of verifying the authenticity of media content, comprising the steps of:

- extracting a sequence of authentication bits from said media content by comparing a property of the media content in successive sections of the media content with a threshold,
- receiving a sequence of authentication bits, said received sequence being extracted from an original version of the media content by comparing said property of the media content with a first threshold, and
- declaring the media content authentic if the received sequence of authentication bits matches the extracted sequence of authentication bits,

characterized in that the step of extracting the authentication bits from the media content comprises setting the threshold in dependence upon the received authentication bits such that the probability that an extracted authentication bit matches the corresponding received authentication bit is increased compared with using the first threshold.

2. A method of verifying the authenticity of media content, comprising the steps of:

- extracting a sequence of authentication bits from said media content by comparing a property of the media content in successive sections of the media content with a threshold,
- receiving a sequence of authentication bits representing an original version of the media content, and
- declaring the media content authentic if the received sequence of authentication bits matches the extracted sequence of authentication bits,

characterized in that the step of extracting the authentication bits from the media content comprises controlling the threshold in dependence upon the received

authentication bits such that the probability that an extracted authentication bit matches the corresponding received authentication bit is relatively high.

3. A method as claimed in claim 1, further comprising a second phase in which the step of extracting is repeated using the first threshold.

4. A method as claimed in claim 1, comprising further phases in which the step of extracting is repeated using a threshold being controlled in dependence upon the distance between the section for which the authentication bit is extracted and sections for which it has been found that the authentication bits do not match with the received authentication bits.

5. A method and arrangement for adjusting the operating point (or, equivalently, the decision boundary, or prior probabilities) according to context information as given, for example, by a neighboring decision

6. Application of the idea within a hypothesis test decision framework (as given in the example implementation)

7. Application of the idea in multimedia (image/video/audio) authentication decisions

8. Adjustment of operating point (or, equivalently, the decision boundary, or prior probabilities) in multimedia authentication decisions according to context information as given, for example, by proximity to areas already determined as tampered

ABSTRACT

The ability to authenticate images captured by a security camera, and localise any tampered areas, will increase the value of these images as evidence in a court of law. This paper outlines the challenges in security camera video authentication, and discusses the reasons why fingerprinting, a robust type of digital signature, provides a solution preferable to semi-fragile watermarking.

Typically, the suspect image is divided into blocks. For each block, an authentication bit is generated by computing some property of the image content and then tresholding said property to give a '0' or '1'. The authentication bits of the suspect image are compared with those of the original image. If there is a mismatch, and the content has indeed been tampered, this is called a *detection*. A mismatch due to allowable operations (e.g. compression) is called a *false alarm*. A so-called ROC curve (Receiver Operating Characteristic) gives the relation between detection probability and false alarm probability. The threshold used to determine the authentication bits represents an *operation point* on the ROC curve.

In accordance with the invention, an operation point corresponding to a low false alarm probability is initially chosen. Mismatching authentication bits may then be considered to have indeed been caused by tampering. This justifies the conclusion that the image has been tampered. However, a low false alarm rate also corresponds to low detection probability: many tampered blocks will not be detected. This leads to a patchy detection of the tampered image area. In order to more precisely identify the tampered image area, the authentication decisions are repeated for neighboring blocks, using a different operation point. This continues until no further tampered blocks are found.

Fig.: none